Stoch Environ Res Risk Assess (2008) 22:315–324
DOI 10.1007/s00477-007-0117-2

ORIGINAL PAPER

Analyzing spatial ecological data using linear regression and wavelet analysis

Gudrun Carl · Ingolf Kühn

Published online: 16 March 2007

Abstract Spatial (two-dimensional) distributions in ecology are often influenced by spatial autocorrelation. In standard regression models, however, observations are assumed to be statistically independent. In this paper we present an alternative to other methods that allow for autocorrelation. We show that the theory of wavelets provides an efficient method to remove autocorrelations in regression models using data sampled on a regular grid. Wavelets are particularly suitable for data analysis without any prior knowledge of the underlying correlation structure. We illustrate our new method, called wavelet-revised model, by applying it to multiple regression for both normal linear models and logistic regression. Results are presented for computationally simulated data and real ecological data (distribution of species richness and distribution of the plant species *Dianthus carthusianorum* throughout Germany). These results are compared to those of generalized linear models and models based on generalized estimating equations. We recommend wavelet-revised models, in particular, as a method for logistic regression using large datasets.

Keywords Binary data · Distribution data · Multiple regression · Normal data · Spatial autocorrelation

G. Carl · I. Kühn (✉)
Department Community Ecology (BZF),
UFZ-Helmholtz Centre for Environmental Research,
Theodor-Lieser-Strasse 4,
06120 Halle, Germany
e-mail: ingolf.kuehn@ufz.de

G. Carl
e-mail: gudrun.carl@ufz.de

G. Carl · I. Kühn
Virtual Institute Macroecology,
Theodor-Lieser-Strasse 4, 06120 Halle, Germany

1 Introduction

Many ecological studies are based on a statistical analysis of data sampled in a spatial, i.e., two-dimensional context. Such spatial data are a challenge in ecology because they often display so-called spatial autocorrelation (Cressie 1993; Legendre 1993; Lichstein et al. 2002), i.e., adjacent data points are more likely to be similar than distant ones. Unfortunately, standard methods like generalized linear models (GLM) may yield wrong results, if spatial autocorrelation is ignored. Simulation studies (e.g., Anselin and Bera 1998; Lennon 2000) have demonstrated that model parameter estimates may be wrong due to this spatial autocorrelation. Therefore, spatial data often require the application of ''new'' methods.

Wavelets seem to be a relatively unemployed tool in ecological application (Dale et al. 2002). However, wavelet analysis has been proved to be a suitable mean to quantify spatial structure as a function of both scale and position (Bradshaw and Spies 1992; Dale 1999; Xiangcheng et al. 2005). These authors used one-dimensional wavelets to describe spatial pattern in plant communities, though not to remove spatial autocorrelation. Keitt and Urban (2005) applied wavelet transforms to multiple linear regressions. They found that different environmental variables show up as good predictors at different scales. However, they applied wavelets in one dimension only. Moreover, only linear models with normally distributed response variables were studied. Concerning logistic regression, the authors posed the inspiring question ''whether binary wavelet-like transforms can be defined for presence-absence data''.

Since wavelets can split up signals into smooth and noisy components at different scales there is evidence to suggest that a wavelet approach will be able to provide

insights in properties of observations like autocorrelation. The aim of this paper is to present a wavelet analysis to account for spatial non-independence in multiple regressions for both normal linear models and logistic regression, i.e., the distribution of response variables can be normal or binomial (which is typical for the distribution of species, e.g., habitat modelling).

To our knowledge, wavelets have never been used to remove spatial autocorrelation in logistic regressions. Fadili and Bullmore (2001) provided a wavelet-based method for linear regressions in the context of correlated errors, but only for normal linear models. They applied their method to neurophysiological time series. Meyer (2003) showed for such kind of data that a systematic baseline drift belongs to a subspace spanned by specific wavelets. In time series analysis local differencing has become a standard tool to remove autocorrelation (Box and Jenkins 1976). The wavelet transform can provide an alternative to such traditional techniques. The main statistical use of wavelets, however, has been in nonparametric regression (Nason and Silverman 1995) which is a quite different task.

In real vegetation and large-scale analysis correlations are performed across spatially (i.e., two-dimensionally) organized data. Therefore, we perform wavelet analysis for two dimensions. The technique is illustrated by its application to both simulated datasets and real large-scale spatial datasets of plant species distributions in Germany. To demonstrate the effectiveness of our method we compare our results to those of generalized estimating equations (GEE; Liang and Zeger 1986).

2 Methods

2.1 Wavelet decomposition

The advantage of wavelets is clearly to find in the separation of the original signal into several scales. Different properties of the signal are visible at different scales (Louis et al. 1994). We want to use this property of wavelets to remove the effect of autocorrelation in signals or spatial ecological data. To better comprehend our approach, we will give the following succinct review on wavelets.

Wavelets come in families. First we consider mother wavelet ψ and father wavelet ϕ. The mother wavelet integrates to zero, and the father wavelet integrates to one, i.e. $\int \psi(x)\mathrm{d}x = 0 \quad \text{and} \quad \int \phi(x)\mathrm{d}x = 1$.

The mother wavelets are used to describe the detail and high-frequency parts of given data, whereas the father wavelets are used to describe the smooth and low-frequency parts. The oldest and simplest example of a function ψ is the Haar function

$$\psi(x) = \begin{cases} 1 & 0 \le x < 1/2 \\ -1 & 1/2 \le x < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

However, wavelets do not need to have an analytic form in general.

The orthogonal wavelet series approximation (Bruce and Gao 1996, p. 14; Shumway and Stoffer 2000) for a function $f(x)$ is given by

$$f(x) = \sum_{j=1}^{J} \sum_{k} d_{j,k}\psi_{j,k}(x) + \sum_{k} s_{J,k}\phi_{J,k}(x), \tag{2}$$

where the functions $\psi_{j,k}(x)$ and $\phi_{J,k}(x)$ are given orthogonal wavelet functions and the coefficients $d_{1,k}$, ...,$d_{J,k}$, $s_{J,k}$ are the corresponding wavelet transform coefficients. The functions $\psi_{j,k}(x)$ and $\phi_{J,k}(x)$ are wavelets generated from mother and father wavelets ψ and ϕ, respectively, by scaling and translation:

$$\begin{aligned} \psi_{j,k}(x) &= 2^{-j/2}\psi\left(\frac{x - 2^j k}{2^j}\right), \\ \phi_{j,k}(x) &= 2^{-j/2}\phi\left(\frac{x - 2^j k}{2^j}\right). \end{aligned} \tag{3}$$

The wavelet transform calculates the wavelet transform coefficients $d_{1,k}$, ..., $d_{J,k}$, $s_{J,k}$ of the wavelet expansion (2). High magnitude of a wavelet transform coefficient means high similarity between the function $f(x)$ and the corresponding wavelet. Scaling and translation (3) produce wavelets to different levels j and shifts k, i.e., dilations and locations. The information of function $f(x)$ is completely codable by a minimal set of wavelets if the used wavelet family is an orthogonal one. Several orthogonal wavelet families are known. One of them is the ''haar'' family. The families are available, e.g., in the R package *waveslim* (Whitcher 2005).

Due to the character of wavelets, the coefficients $d_{j,k}$ (coefficients d for short) of Eq. (2) are able to represent the detail parts of the shape of function $f(x)$, while the coefficients $s_{J,k}$ (coefficients s for short) are able to represent its smooth part. In general, detail coefficients d have to be calculated at different levels j in contrast to smooth coefficients s . Thus, changes in magnitude of f are captured in detail coefficients for different levels $j = 1, ..., J$. Parts of f which are not included so far are mapped in smooth coefficients s of last level J. Note that higher dilation levels mean lower resolution. The level of lowest resolution J can arbitrarily be chosen. Since the full information of f is captured in the coefficients d and s, it is possible to reconstruct the original function f by application of the inverse wavelet transform.

This inverse transform can also be performed for the following single sums

$$f^j(x) = \sum_k d_{j,k}\psi_{j,k}(x) \quad \text{for} \quad j = 1, \ldots, J$$

$$f^{J+1}(x) = \sum_k s_{J,k}\phi_{J,k}(x)$$

In this way one can obtain the $J + 1$ orthogonal components f^i which add up to f in good approximation:

$$f(x) = \sum_{i=1}^{J+1} f^i(x). \tag{4}$$

This approach is called multiresolution analysis because different resolutions of the function are described by the terms at different levels. The last term of the series of $J + 1$ orthogonal components corresponds to the smooth part of the function. For a more detailed treatment of the theory, we refer to Daubechies (1992).

One of the fields where wavelets are applied is the so-called noise removal of signals. In principle, this is done by taking the wavelet transform of the signal, keeping only coefficients which represent the smooth part of the function, and performing the inverse transform. In contrast the aim of this paper is to remove such coefficients which cause autocorrelation in the original data. Here we deal with datasets as they appear in statistical samples, especially in linear regressions. To this end we use the wavelet transform in a special form, the so-called discrete wavelet transform, which calculates the coefficients for a finite set of discrete data.

2.2 Normal linear models

Now we concentrate on multiple regression of the following normal linear model (LM)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{5}$$

where $\mathbf{y}$ is the vector of response variables, $\mathbf{X}$ is the design matrix of predictors, $\boldsymbol{\beta}$ is the vector of regression parameters, and $\boldsymbol{\varepsilon}$ is the vector of errors. The ordinary least squares estimator of $\boldsymbol{\beta}$ is given by

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{6}$$

where $\mathbf{X}'$ is the transposed matrix of $\mathbf{X}$.

If we consider the case that a wavelet transform is applied to the normal linear model, we can expect that the regression parameters will remain unchanged. This is due to the fact that the wavelet transform T which maps a vector of data to a vector of wavelet transform coefficients is a linear operator. Its application to Eq. (5) yields

$$T(\mathbf{y}) = T(\mathbf{X}\boldsymbol{\beta}) + T(\boldsymbol{\varepsilon}) = (T(\mathbf{X}))\boldsymbol{\beta} + T(\boldsymbol{\varepsilon}). \tag{7}$$

The transform $T(\mathbf{X})$ of the design matrix $\mathbf{X}$ is given by

$$T(\mathbf{X}) = (T(\mathbf{1}), T(\mathbf{x}_1), \ldots, T(\mathbf{x}_p)),$$

where $\boldsymbol{X} = (\mathbf{1}, \boldsymbol{x_1}, \ldots, \boldsymbol{x_p})$.

Here $\mathbf{x}_1, \ldots, \mathbf{x}_p$ are the predictors and $\mathbf{1}$ means a vector of equal length filled by ones. Then we can use ordinary least squares on the transformed data as well. The solution is

$$\mathbf{b} = ((T(\mathbf{X}))'T(\mathbf{X}))^{-1}(T(\mathbf{X}))'T(\mathbf{y}). \tag{8}$$

The covariance matrix of $\mathbf{b}$ is $\sigma_T^2((T(\mathbf{X}))'T(\mathbf{X}))^{-1}$ where σ_T^2 is the variance of the transformed errors $T(\boldsymbol{\varepsilon})$.

If the errors $\boldsymbol{\varepsilon}$ in Eq. (5) are correlated standard regression is no longer applicable. Then standard regression also fails in the wavelet domain even though the variance-covariance matrix of $T(\boldsymbol{\varepsilon})$ in (7) is approximately diagonalized. Its diagonal elements, however, are no longer constant.

2.3 Logistic regression model

We are going to specify the Eqs. (5, 6) for application in logistic regression (Hosmer and Lemeshow 2000; Collett 2003). In particular, we want to describe a model with binary responses, i.e., the number of trials equals one. The logit link is given by

$$\text{logit}(\pi_i) = \log((\pi_i/(1 - \pi_i)) = \mathbf{x}_i'\boldsymbol{\beta} = \eta_i \quad i = 1, 2, \ldots, n$$

with the expected value of response $E(y_i) = \pi_i$, the variance $\text{var}(y_i) = \pi_i\,(1 - \pi_i)$, and the sample size n. The solution form for generalized linear models (GLM) is

$$\mathbf{b}^{(m)} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z} \tag{9}$$

where $\mathbf{b}^{(m)}$ is the vector of estimates at the mth iteration. In case of binary responses $\mathbf{z}$ has elements

$$z_i = \eta_i + (y_i - \pi_i)\frac{\partial \eta_i}{\partial \pi_i} = \eta_i + \frac{y_i - \pi_i}{\pi_i(1 - \pi_i)} \tag{10}$$

and the weights are functions of the fitted values $\mathbf{W} = diag\{\pi_i\,(1 - \pi_i)\}$ (Dobson 2002, p. 64; Myers et al. 2002, p. 330). Equation (9) is the least squares estimator of the following linear regression

$$\mathbf{W}^{1/2}\mathbf{z} = \mathbf{W}^{1/2}\mathbf{X}\boldsymbol{\beta}^{(m)} + \boldsymbol{\varepsilon}.$$

It can be transformed by wavelets in a similar way as Eq. (5), which yields

$$T(\mathbf{W}^{1/2}\mathbf{z}) = T(\mathbf{W}^{1/2}\mathbf{X})\boldsymbol{\beta}^{(m)} + T(\boldsymbol{\varepsilon}), \tag{11}$$

where $\boldsymbol{\beta}^{(m)}$ remains unchanged. Therefore, another least squares estimator of $\boldsymbol{\beta}^{(m)}$ is given by

$$\mathbf{b}^{(m)} = ((T(\mathbf{W}^{1/2}\mathbf{X}))'T(\mathbf{W}^{1/2}\mathbf{X}))^{-1}(T(\mathbf{W}^{1/2}\mathbf{X}))'T(\mathbf{W}^{1/2}\mathbf{z}). \tag{12}$$

Moreover, Eq. (10) can be transformed as follows

$$T(\mathbf{W}^{1/2}\mathbf{z}) = T(\mathbf{W}^{1/2}\mathbf{X}\mathbf{b}^{(m-1)} + \mathbf{W}^{-1/2}(\mathbf{y} - \boldsymbol{\pi})). \tag{13}$$

The Eqs. (12, 13) solved iteratively yield the same results as Eqs. (9, 10). The asymptotic variance-covariance matrix of $\mathbf{b}$ is given by var $(\mathbf{b}) = ((T(\mathbf{W}^{1/2}\ \mathbf{X}))'T(\mathbf{W}^{1/2}\mathbf{X}))^{-1}$.

Recall that $\mathbf{W}$ is assumed to be diagonal. Only if $\mathbf{W}$ would contain off-diagonal elements, then correlations between responses could be described. But this is not the case here. Therefore, it is desirable to revise data and to remove autocorrelations. We are going to prove that one can achieve this goal by using the concept of wavelet decomposition. Our new method, called wavelet-revised model and introduced in Sect. 2.6, will provide this revision of data.

2.4 Autocorrelation

As we have pointed out above, multiresolution analysis provides an additive decomposition (4) of a given function f. A given vector $\mathbf{f}$ can accordingly be decomposed using discrete wavelet transforms. In this chapter we are going to clarify whether this property of additivity holds for the autocorrelation of these data.

For this purpose we consider the autocovariance for a lag distance that equals one, i.e., the influence of nearest neighbours. This covariance for a vector $\mathbf{f}$ with components f_1, f_2, ... and for neighbourhood as neighbourhood of components is defined by

$$\mathrm{cov}(1) = \sum_n f_n f_{n+1}.$$

Here each component of the vector $\mathbf{f}$ can be decomposed according to the multiresolution analysis (4). This yields a product of sums which can be reordered in the following way

$$f_n f_{n+1} = \left(\sum_j^{J+1} f_n^j\right)\left(\sum_k^{J+1} f_{n+1}^k\right) = \sum_j^{J+1} c_{n,n+1}^j$$

with components

$$c_{n,n+1}^j = f_n^j f_{n+1}^j + \sum_{k=1}^{j-1}(f_n^j f_{n+1}^{j-k} + f_n^{j-k} f_{n+1}^j)$$

which include all across-the-level parts of former resolution levels. By using the components $c_{n,n+1}^j$ the autocovariance for lag distance of one can be rewritten in the following form

$$\mathrm{cov}(1) = \sum_j^{J+1}\sum_n c_{n,n+1}^j = \sum_j^{J+1} \mathrm{cov}^j(1). \tag{14}$$

Equation (14) shows that cov (1) is additive with respect to the level parts. Autocovariances for other lags can accordingly be calculated. Moreover, we obtain the level parts of autocorrelation from the level parts of autocovariance standardized by the total variance.

2.5 The two dimensional approach

The aim of this chapter is to analyze data points which are spatially distributed. For this purpose we use the two-dimensional (2D) wavelet approximation of a function $F(x,y)$ as follows (Bruce and Gao 1996, p. 44):

$$F(x,y) = \sum_{m=1}^{3}\sum_{j=1}^{J}\sum_{k_x,k_y} d_{j,k_x,k_y}^m \Psi_{j,k_x,k_y}^m(x,y) + \sum_{k_x,k_y} s_{J,k_x,k_y}\Phi_{J,k_x,k_y}(x,y). \tag{15}$$

Equation (15) represents an extension of Eq. (2), where m corresponds to different spatial directions (see below). The discrete wavelet transform calculates the coefficients for a finite set of discrete data points in the same way as above. Thus the 2D discrete wavelet transform enables us to transform discrete image data $\mathbf{F}_{p,q}$ into a $p \times q$ matrix of wavelet coefficients. This allows us to analyze 2D data such as a matrix or a geographical pattern of an ecological or an environmental variable. In what follows we are going to describe the different meanings of the wavelet coefficients. The coefficients d represent the detail part, while the coefficients s represent the smooth part of $\mathbf{F}$. The detail coefficients d depend on both level j and spatial direction m, in contrast to smooth coefficients s. Thus changes in magnitude of $\mathbf{F}$ are captured in detail coefficients for different levels $j = 1, ..., J$ and for different directions $m = 1, 2, 3$. These three directions correspond to wavelets which work horizontally, vertically or diagonally as mother wavelets. Remaining parts of $\mathbf{F}$ are mapped to smooth coefficients s of the last level J. For some related remarks, see Müller et al. (2003).

A 2D approach can be applied to both the response variables $\mathbf{y}$ and individual predictors $\mathbf{x}_1$, ..., $\mathbf{x}_p$ of multiple regression models, if the components of these variables occur in a spatial context, e.g., if these components were sampled in a plane. Thus we have to convert these vectors into matrices which reflect the special spatial form. Then the 2D transform can be applied to each matrix built in this way. Finally, we can go back to vectors which allow us to continue as it is usual in linear regression models.

2.6 Wavelet-revised models

In the following transform T will denote the wavelet procedure for decorrelation in the broader sense. This removal of autocorrelation is done by taking the wavelet transform of a vector $\mathbf{f}$, and keeping only coefficients which represent the detail parts of the vector. The regression problem can be solved either in the wavelet domain or after back transform.

For instance, consider a vector $\mathbf{f}$ whose smooth part was removed by multiresolution analysis (4). This can be done at different resolution levels J. For 2D wavelet transforms one has to sum, in addition, over three spatial directions m. Thus, the reduced vector $\mathbf{f}^{\text{part}}$ can be written as

$$\mathbf{f}^{\text{part}} = \sum_{i}^{3J} \mathbf{f}^{i},$$

whereas the total vector is

$$\mathbf{f}^{total} = \sum_{i}^{3J+1} \mathbf{f}^{i}.$$

Thus the wavelet procedure used here is defined by T: $\mathbf{f}^{\text{total}} \rightarrow \mathbf{f}^{\text{part}}$. To a good approximation, transform T provides a method to remove autocorrelation. It enables us to calculate regression parameters under the assumptions pointed out above. This can be done by Eq. (8) for normal responses or iteratively by Eqs. (12, 13) for binary responses. Therefore, if observations are originally autocorrelated, these equations could improve parameter estimates compared to LM (6) and GLM (9, 10), respectively. The Eqs. (8, 12, 13) combined with T: $\mathbf{f}^{\text{total}} \rightarrow \mathbf{f}^{\text{part}}$ form our new method, called wavelet-revised model (WRM) because transform T provides a revision of data with regard to autocorrelation.

To see if our procedure effectively removed spatial autocorrelation at a certain level, it is necessary to test residual autocorrelation in the original domain. For this we used Moran's I coefficient (see e.g., Dale et al. 2002; Lichstein et al. 2002) which describes a radial autocorrelation for 2D data. The autocorrelation of vector $\mathbf{f}^{\text{part}}$ with elements f_1^{part}, f_2^{part}, ... can be written in the following form:

$$cor^{\text{part}}(d) = \frac{\frac{1}{s}\sum_k \sum_l w_{kl}(f_k^{\text{part}} - f_{\text{mean}}^{\text{part}})(f_l^{\text{part}} - f_{\text{mean}}^{\text{part}})}{\frac{1}{n}\sum_k (f_k^{\text{total}} - f_{\text{mean}}^{\text{total}})^2} \qquad (16)$$

Here one has to introduce ''lag distance'' intervals d for the spatial structure under consideration. The factor w_{kl} is a weight that equals one if the distance of the variables f_k^{part} and f_l^{part} belongs to this interval d and which equals zero otherwise. The factor s is the sum of weights for a given interval d and n is the length of the vectors.

3 Application

3.1 Implementing WRM

Our computations are based on software packages in the computer language R (R Development Core Team 2004). The tools for calculating wavelet transforms are available in package *waveslim* (Whitcher 2005). We used either the functions *dwt.2d* and *idwt.2d* for discrete wavelet transform and inverse discrete wavelet transform, respectively, or the function *mra.2d* for multiresoluion analysis. These functions offer various wavelets. We used the ''haar'' family (1) not only because it is the simplest kind of wavelets but also for reasons that we shall see later. Furthermore, the autocorrelation (16) of vector $\mathbf{f}^{\text{part}}$ has to be calculated. For this purpose we wrote an R-code using fast Fourier transform and convolutions.

Because of the truncation to finite sets in discrete wavelet transforms it is necessary to give boundary treatment rules. In the 2D discrete wavelet transform type *periodic* is implemented for boundary conditions. This causes a restriction on the sample size. The number of rows and columns must be divisible by 2^j in order to perform dilation and location of wavelets as described in Eq. (3). In general, one wishes to analyze samples of arbitrary size though. For this reason we decided to pad data with zeros until a quadratic matrix of required size is reached.

Generally, data padded with zeros and decomposed by multiresolution analysis have the inconvenient property slightly to lose the contour. This phenomenon occurs, in particular, if smooth parts of data are analyzed by means of wavelets of wide and smooth shape. For this reason we use ''haar'' wavelets which have a compact and square-edged shape. ''Haar'' wavelets are useful in the detection of edges and gradients (Bradshaw and Spies 1992) in contrast to smooth ones. Moreover, our data processing task is different. To find uncorrelated data we need to choose

detail parts and to remove smooth ones. If we perform this task, then autocorrelations calculated via fully mapped residuals will hardly to be distinguished from those calculated via residuals truncated at the accurate original contour of the dataset. Therefore, we provide residuals and autocorrelations of the last kind.

To show the effectiveness, we compare several wavelet-revised models (WRM) to generalized linear models (GLM) and generalized estimating equations (GEE) for both binary data and normal data. The tools for calculating GEE are available in package *gee* (Carey et al. 2002) with function *gee* (Zeger and Liang 1986; Diggle et al. 1995) and in package *geepack* (Yan 2004) with function *geese* (Yan 2002; Yan and Fine 2004). These functions offer various correlation structures. We know as a result of tests and previous analyses that fixed and user defined correlation structures work best in the cases considered here.

3.2 Simulation

Simulations were performed to check the models regarding autocorrelation effects. For this purpose regular grids were generated. The number of grid cells is 34×34 and the cells were assumed to be square. Values for two normally distributed predictors were randomly generated, and linearly combined using specified parameters (intercept and two slopes). In addition, normally distributed errors were randomly generated. Both the vector of errors and the vectors of the predictors were multiplied by the Cholesky decomposition of a variance–covariance matrix. This procedure creates correlated normal random errors and predictors. Finally, we are able to calculate correlated responses. On one hand normal responses are given as the sum of linear component and correlated errors. On the other hand the following steps transform these correlated normal variables into correlated binary outcomes: (1) scale to get the standard normal distribution, (2) transform by their cumulative distribution function to get a uniform distribution, and (3) use the inverse transform method to get binomial outcomes.

The correlation matrix includes specified spatial autocorrelation depending on the distances between the points of measurements (e.g., centre points of grid cells). In our case this correlation is assumed to be equal for each pair of equal distance. In this way we have introduced an isotropic spatial autocorrelation structure by an exponential function. Note that two scale parameters are necessary for binary response data. The first one ensures the correct fitting in case of uncorrelated errors. In order to check the fit GLM regressions for 1,000 simulated datasets are performed. The second scale parameter has to preserve the specified error variance when the correlation is incorporated.

3.3 Application to a simulated grid

Figure 1 shows the remaining parts of residual autocorrelation (16) for an average of 50 generated datasets of both normal and binomial (binary) distribution. Autocorrelation is reduced by removing of smooth wavelet coefficients at different levels. As it can be seen in Fig. 1 the degree of correlation removal depends on the resolution level. The level ranges from six (coarsest resolution) to one (finest resolution). If the level is large, then smooth wavelet coefficients are reduced only at a coarse scale and autocorrelation is hardly reduced compared to the full model (GLM). By contrast, if the level is small, then smooth wavelet coefficients are reduced up to the finer scales and autocorrelation disappears. Our approach obtained by visual inspection of these trials is to select level 1 for normal data. For binary data level 2 corrects correlation well, while level 1 slightly overestimates the nearest neighbour correlation. Thus we use level 2 in further applications for binary data. Note that in case of binomial distribution the iterations related to Eqs. (12, 13) will destroy, in general, the additivity of the level parts. However, our calculations have shown that the differences are negligible.

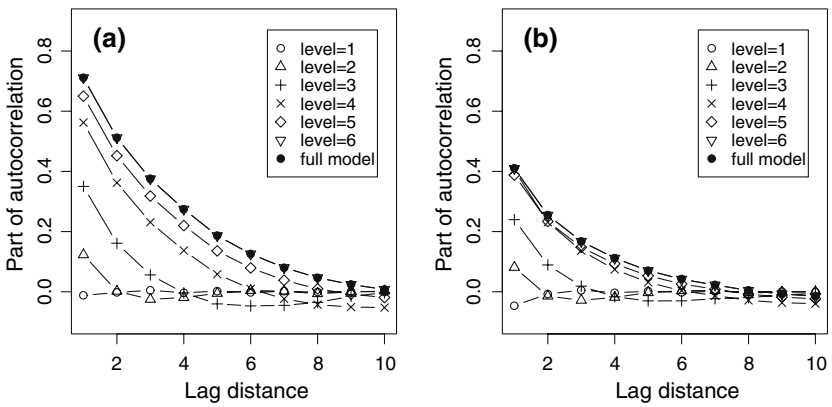

Fig. 1 Residual autocorrelation for an average of 50 randomly generated 34×34 datasets of **a** normal distribution and **b** binary distribution. Autocorrelation is reduced by removing of smooth wavelet coefficients at different levels

Table 1 Means and variances for estimated regression parameters (Intercept = b_0, slope of first predictor = b_1, slope of second predictor = b_2) calculated from 1000 randomly simulated 34 × 34 datasets for different distributions and compared for different methods

β	mean(b_0) –1	mean(b_1) 3	mean(b_2) –2	var(b_0)	var(b_1)	var(b_2)
Normal response						
GLM	–0.9925	2.9994	–2.0032	0.0591	0.0366	0.0339
GEE	–0.9940	2.9993	–2.0015	0.0246	0.0111	0.0108
WRM	–0.9920	2.9977	–2.0016	0.0351	0.0127	0.0131
Binary response						
GLM	–1.0028	3.0363	–2.0320	0.2090	0.1524	0.1286
GEE	–0.9983	3.0203	–2.0216	0.1208	0.0940	0.0761
WRM	–0.9964	3.0114	–2.0228	0.1674	0.1098	0.0879

Next we will discuss the accuracy and efficiency of parameter estimates. In Table 1 we present results for 1,000 randomly simulated datasets of sample size 34 × 34 for both normal and binomial distribution. The means and variances for regression parameters are given for different methods: (1) GLM, (2) GEE as above-mentioned method for comparison, (3) wavelet-revised models (WRM) using ''haar'' wavelets and smooth coefficients removal. All simulated datasets were created with equal regression parameters. The real parameters are $\beta_0 = -1$, $\beta_1 = 3$, and $\beta_2 = -2$. As it can be seen all means fit very well. However, we recognize differences in the variances. The maximum values for variances and, therefore, the lowest efficiency for parameter estimates can clearly be found for GLM. Variances decrease for GEE and WRM. Thus these models provide an efficiency gain compared to GLM. It is of course expected that WRM variances should be somewhat greater than GEE variances. This is due to the prior knowledge about the form of the error covariance matrix. It is completely involved in GEE. Note that there is no need for such specifications in WRM.

Figure 2 gives the type 1 error calibration curves of the methods (Fadili and Bullmore 2001). For this purpose we generated 500 datasets for both distributions. Their autocorrelated responses are independent on a certain autocorrelated predictor. The test results for the null hypothesis H_0: $\beta = 0$ for probabilities of type 1 error α are summarized. All plots in Fig. 2 show the observed number of positive tests per 500 realizations versus the expected number per 500 realizations in the full range of α. Methods work best when the observed number equals the predicted one, i.e. when the calibration curve coincides with the line of identity given as straight line in all plots. It is shown that GLM overestimates the true type 1 error, whereas both

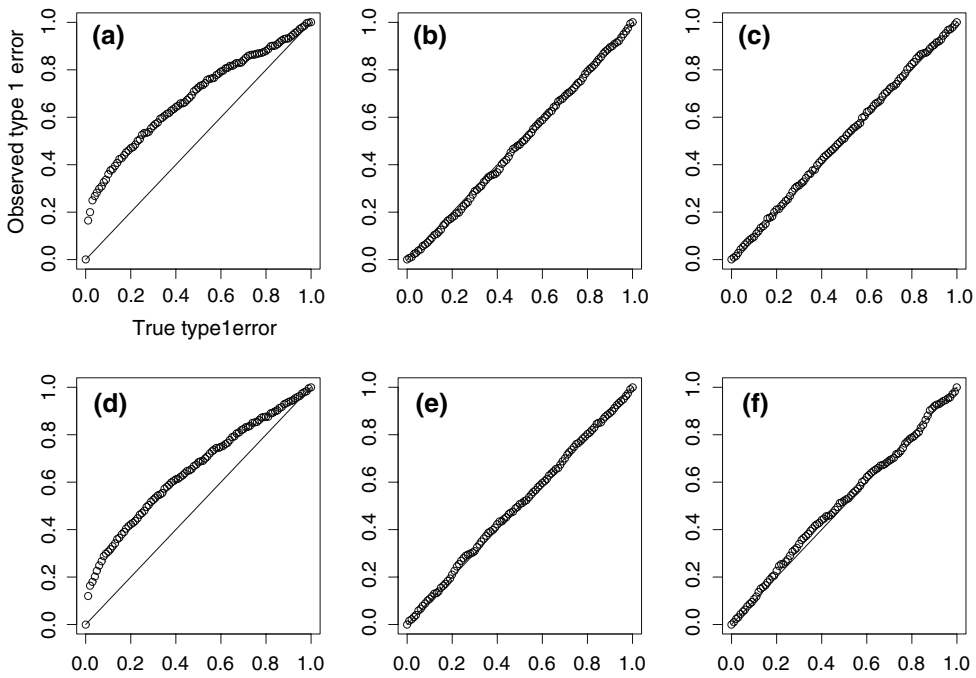

Fig. 2 Type 1 error calibration curves for simulated datasets of **a–c** normal distribution and **d–f** binary distribution. The compared methods are **a**, **d** GLM, **b**, **e** GEE, **c**, **f** WRM

GEE and WRM yield very good error calibration curves for normal and binomial distributions.

3.4 Application to the flora of Germany

In this section we apply the wavelet-revised methods to real macroecological datasets. We relate environmental variables to plant species distribution in Germany. Information on species distribution is available from FLORKART (see http://www.floraweb.de) which contains species location in a grid of 2,995 grid cells. The cells of this lattice are 10′ longitude × 6′ latitude, i.e., about 11×11 km^2, and therefore almost square cells. We selected species data for two regression models. They differ in the response variable.

1. A dataset has been built with the normally distributed number of all plant species found per grid cell. Figure 3a shows the distribution of species richness in Germany with green for species poor cells and red for species rich cells.
2. A dataset for logistic regression has been chosen with binary distributed responses for presence/absence of the plant species *Dianthus carthusianorum* whose ecological behaviour is well known. *D. carthusianorum* is a species typical for nutrient poor dry and semi-dry grasslands on calcareous as well as siliceous soils mostly in mountainous areas but occasionally on sandy grasslands in the lowlands. The distribution of *D. carthusianorum* is given in Fig. 3e (green: presence, red: absence).

Both datasets have the same predictors. We only choose two environmental variables: (1) mean annual temperature based on a 1 km^2 grid scale was provided by the ''Deutscher Wetterdienst, Department Klima und Umwelt''. Recording period for temperature data was 1951–1980. (2) Averages altitude (in 100 m units) per grid cell was calculated after the ARCDeutschland500 dataset, scale 1:500,000, provided by ESRI.

Figure 4 shows the parts of autocorrelation for residuals as described for Fig. 1. Figure 4a presents the results for normally distributed numbers of all plant species in Germany. Here level 1 provides the best reduction of correlation. Figure 4b presents the results for binary distributed responses for presence/absence of the special plant species *Dianthus carthusianorum*. According to our experience with simulated data level 2 is chosen for decorrelation since level 1 leads to an overestimation of nearest neighbour correlation.

Hence, wavelet-revised regression was carried out as described above using ''haar'' wavelets, and smooth coefficients removal at the contemplated levels. Results for regression parameters for normal data of plant species richness in Germany and for binary data for presence/absence of *Dianthus carthusianorum* are given in Table 2. Here WRM can be compared to GEE and GLM. In fact, WRM and GEE provide more or less similar values for

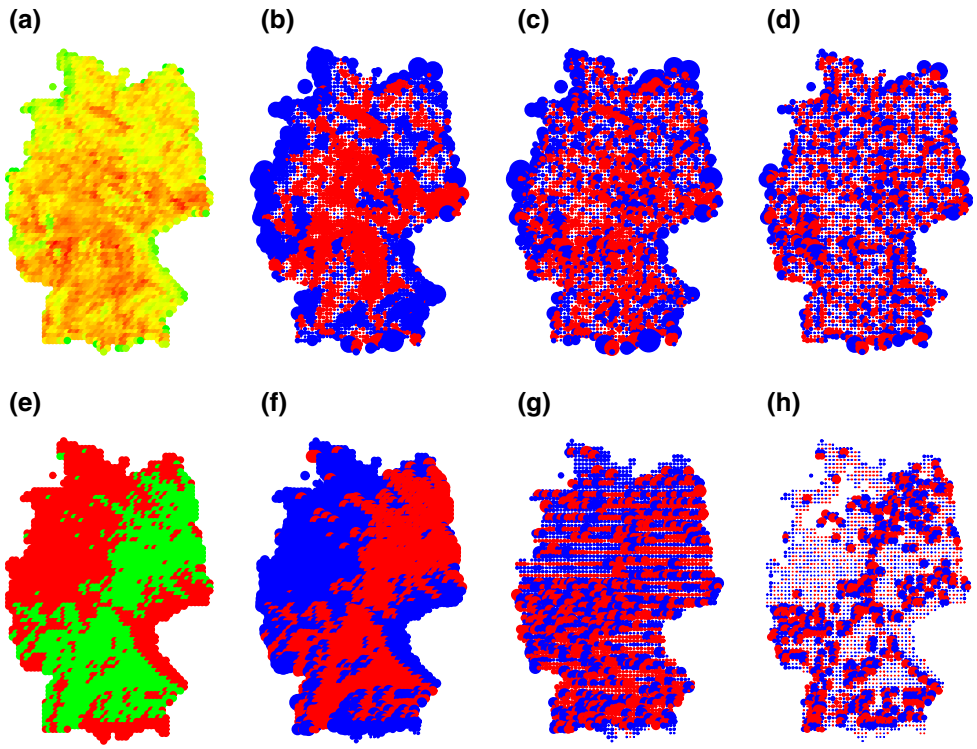

Fig. 3 Distribution of data and residuals across Germany for **a** normal data for plant species richness with **b** corresponding GLM residuals **c** corresponding GEE residuals **d** corresponding WRM residuals, and **e** binary data for presence/absence of *Dianthus carthusianorum* with **f** corresponding GLM residuals, **g** corresponding GEE residuals, **h** corresponding WRM residuals

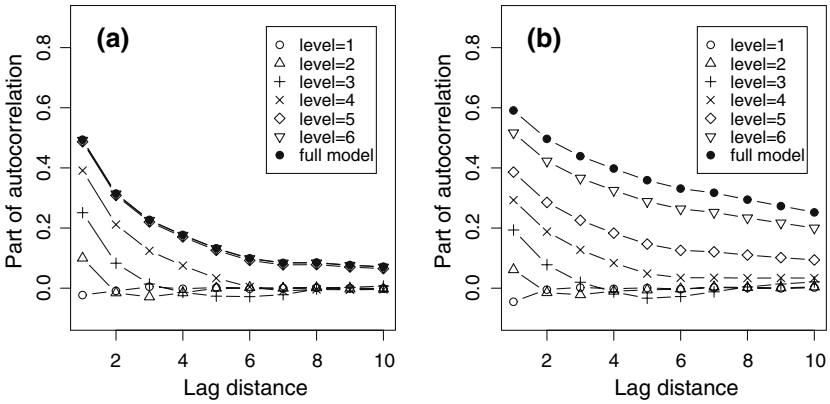

Fig. 4 Residual autocorrelation **a** for normal data of plant species richness in Germany and **b** for binary data for presence/absence of *Dianthus carthusianorum* in Germany. Autocorrelation is reduced by removing of smooth wavelet coefficients at different levels

slopes in normal distribution. For binary data of this sample size it is difficult to find a proper GEE approach. Here we present the results of an approximation that partly neglects correlations. However, we are able to discuss the results on the basis of our geographical and biological knowledge. For binary data for presence/absence of *D. carthusianorum*, see Fig. 3e. This plant species can only rarely be found in the western lowlands or on high mountains in the south of Germany. Hence, its presence/absence data should not be positive correlated with predictor Altitude. In fact, the GLM results are corrected by WRM in this way. The corresponding parameter is reduced from 0.26 to –0.06. Moreover, it is no longer significant as seen by its *p*-value.

Furthermore, the spatial distribution of residuals is given in Fig. 3. Here pixel sizes indicate sizes of residuals, the two colours red and blue represent the signs of residuals, and areas of equal colour indicate autocorrelation. For WRM residuals (see Fig. 3d, h) areas of equal colour are essentially reduced compared to GLM residuals (Fig. 3b, f). Note that the white areas in Fig. 3h correspond to residuals that equals zero. They do not contribute to autocorrelation.

Table 2 Results for estimated regression parameters *b* and their *p*-values comparing different methods having normal data of plant species richness in Germany and binary data for presence/absence of *Dianthus carthusianorum* in Germany

Predictor	GLM		GEE		WRM	
	b	*p*	*b*	*p*	*b*	*p*
Plant species richness (normal)						
(Intercept)	–57.2	0.057	203.3	<0.001	57.9	0.059
Altitude	31.2	<0.001	11.0	<0.001	14.1	<0.001
Temperature	78.7	<0.001	50.5	<0.001	47.4	<0.001
Dianthus carthusia-norum (binary)						
(Intercept)	–7.62	<0.001	–5.84	<0.001	–5.13	<0.001
Altitude	0.26	<0.001	0.11	0.002	–0.06	0.434
Temperature	0.80	<0.001	0.63	<0.001	0.49	<0.001

4 Conclusion

We presented a strategy for correcting data with respect to autocorrelation and demonstrated its application to normal and logistic regression models. This strategy was based on discrete wavelet transforms and was carried out as 2D analysis. We used ''haar'' wavelets which are useful in the detection of edges and gradients (Bradshaw and Spies 1992). All datasets exhibited a considerable amount of spatial autocorrelation. Thus, ordinary least square regressions or logistic regressions could result in wrong parameter estimates (Anselin and Bera 1998; Lennon 2000) as the basic assumption of independence of residuals is violated. We compared the results of generalized linear models (GLM) with those of generalized estimating equations (GEE), which are known to successfully correct autocorrelation effects, and the newly developed method of wavelet-revised regression. Both latter methods yielded comparable results and when applying real data, they differed markedly from non-spatial regression models. Wavelets thus provide a powerful method for removing autocorrelation in spatial (e.g. 2D) datasets. Wavelets still seem to be a relatively unemployed tool in ecological application. In particular, this is the first study that we are aware of which corrects spatial autocorrelation in logistic regression models.

To examine the behaviour of our wavelet-revised model (WRM) in detail we calculated regression parameters and their *p*-values for randomly simulated datasets with correlated residuals and predictors. WRM is shown to be more efficient and to give perfect type 1 error calibration curves compared to GLM. Furthermore, we applied WRM to real macroecological datasets. WRM and GEE reduced the slopes compared to GLM. For the example of binary distribution the values of slopes are quite different for GLM and WRM. Basing the findings not only on statistics but also on prior knowledge in ecology and geography, we showed that the WRM results are more plausible than the GLM results.

Using wavelets one is able stepwise to reduce the autocorrelation of regression residuals as measured by a modified Moran's *I* equation. The autocorrelation can be reduced to nearly zero. This can be done for variables in linear multiple regression for both normally and binary distributed responses. We therefore recommend wavelet-revised models in this sense, in particular, as one alternative method for the study of binary data and large datasets, because: (1) WRM effectively removes spatial autocorrelation, (2) it is a computationally very fast and efficient procedure, and (3) it is often easier applied than analogous proper GEE methods. Our recommendation is justified due to our experience with tests of numerous other datasets.

Acknowledgements Gudrun Carl acknowledges a stipend from the federal state ''Sachsen-Anhalt'', Ministry of Education and Cultural Affairs. We acknowledge additional support by Integrated Project ''ALARM: Assessing LArge scale environmental Risks with tested Methods'' (GOCE-CT-2003-506675) (Settele et al. 2005) from European Commission within Framework Programme 6.

References

Anselin L, Bera AK (1998) Spatial dependence in linear regression models with an introduction to spatial econometrics. In: Ullah A, Giles DEA (eds) Handbook of applied econometric statistics. Marcel Dekker, New York, pp 237–289

Box GE, Jenkins GM (1976) Time series analysis: forecasting and control. Holden Day, San Francisco

Bradshaw GA, Spies TA (1992) Characterizing canopy gap structure in forests using wavelet analysis. J Ecol 80:205–215

Bruce A, Gao HY (1996) Applied wavelet analysis with S-plus. Springer, Heidelberg

Carey VJ, ported to R by Lumley T (versions 3.13 and 4.4), Ripley B (version 4.13). (2002). gee: generalized estimation equation solver. R package version 4.13-10

Collett D (2003) Modelling binary data, 2nd edn. Chapman & Hall, London

Cressie NAC (1993) Statistics for spatial data. Wiley, Cambridge

Dale MRT (1999) Spatial pattern analysis in plant ecology. Cambridge University Press, Cambridge

Dale MRT, Dixon P, Fortin MJ, Legendre P, Myers DE, Rosenberg MS (2002) Conceptual and mathematical relationships among methods for spatial analysis. Ecography 25:558–577

Daubechies I (1992) Ten lectures on wavelets. CSBM-NSF series application mathematics, vol. 61. SIAM Publication, Philadelphia

Diggle PJ, Liang KY, Zeger SL (1995) Analysis of longitudinal data. Clarendon, Oxford

Dobson AJ (2002) An introduction to generalized linear models, 2nd edn. Chapman & Hall, London

Fadili MJ, Bullmore ET (2001) Wavelet-Generalized Least Squares: A new BLU estimator of linear regression models with 1/f errors. NeuroImage 15:217–232

Hosmer DW, Lemeshow S (2000) Applied logistic regression, 2nd edn. Wiley, New York

Keitt TH, Urban DL (2005) Scale-specific inference using wavelets. Ecology 86(9):2497–2504

Legendre P (1993) Spatial autocorrelation—trouble or new paradigm. Ecology 74:1659–1673

Lennon JT (2000) Red-shifts and red herrings in geographical ecology. Ecography 23:101–113

Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrica 73:13–22

Lichstein JW, Simons TR, Shriner SA, Franzreb KE (2002) Spatial autocorrelation and autoregressive models in ecology. Ecol Monogr 72(3):445–463

Louis AK, Maaß P, Rieder A (1994) Wavelets: Theorie und Anwendungen. Teubner, Stuttgart

Meyer FG (2003) Wavelet based estimation of a semi parametric generalized linear model of fMRI time-series. IEEE Trans Med Imaging 22(3):315–322

Müller K, Lohmann G, Zysset S, von Cramon DY (2003) Wavelet statistics of functional MRI data and the general linear model. J Magn Reson Imaging 17:20–30

Myers RH, Montgomery DC, Vining GG (2002) Generalized linear models. Wiley, New York

Nason GP, Silverman BW (1995) The stationary wavelet transform and some statistical applications. In: Antoniadis A, Oppenheim G (eds) Wavelets and statistics. Springer, Heidelberg, pp 281–299

R Development Core Team (2004) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, ISBN 3-900051-07-0 http://www.R-project.org

Settele J, Hammen V, Hulme PE, Karlson U, Klotz S, Kotarac M, Kunin WE, Marion G, O'Connor M, Petanidou T, Peterseon K, Potts S, Pritchard H, Pyšek P, Rounsevell M, Spangenberg J, Steffan-Dewenter I, Sykes MT, Vighi M, Zobel M, Kühn I (2005) ALARM: assessing large scale environmental risks for biodiversity with tested methods. GAIA Ecol Perspect Sci Humanit Econ 14:69–72

Shumway RH, Stoffer DS (2000) Time series analysis and its applications. Springer texts in statistics. Springer, Heidelberg

Whitcher B (2005) Waveslim: basic wavelet routines for one-, two- and three-dimensional signal processing. R package version 1.5 http://www.image.ucar.edu/staff/whitcher/, http://www.image.ucar.edu/staff/whitcher/book/

Xiangcheng M, Haibao R, Zisheng O, Wie W, Keping M (2005) The use of the Mexican Hat and the Morlet wavelets for detection of ecological patterns. Plant Ecol 179:1–19

Yan J (2002) Geepack:yet another package for generalized estimating equations. R News 2(3):12–14

Yan J (2004) Geepack: generalized estimating equation package. R package version 0.2-10

Yan J, Fine J (2004) Estimating equations for association structures. Stat Med 23:859–874

Zeger SL, Liang KY (1986) Longitudinal data analysis for discrete and continuous outcomes. Biometrics 42:121–130